JIVFww

Sergeev S, Diakova I, Nadirashvili L. Neural networks pipeline for quality management in IVF laboratory. *Journal of IVF-Worldwide*. 2024;2(4):1-9. doi:10.46989/001c.124947

Original Research Articles

Neural networks pipeline for quality management in IVF laboratory

Sergei Sergeev[1,2a], Iuliia Diakova[3], Lasha Nadirashvili[1]

[1] IVF, Georgian-German Reproductive Center, Tbilisi, Georgia, [2] IVF, IVF and Genetic Center, 105043 Moscow, Russian Federation, [3] IVF, ART-IVF Clinic, 119435 Moscow, Russian Federation

Keywords: Deep neural networks, Kolmogorov-Arnold networks, IVF laboratory data analysis, Clinical pregnancy prediction, Personalized treatment, ART quality management, AIAIVF

https://doi.org/10.46989/001c.124947

Journal of IVF-Worldwide

Vol. 2, Issue 4, 2024

This study introduces a novel neural network-based pipeline for predicting clinical pregnancy rates in IVF treatments, integrating both clinical and laboratory data. We developed a metamodel combining deep neural networks and Kolmogorov-Arnold networks, leveraging their complementary strengths to enhance predictive accuracy and interpretability. The metamodel achieved robust performance metrics after training and fitting on 11500 clinical cases: accuracy = 0.72, AUC = 0.75, F1 score = 0.60, and Matthews Correlation Coefficient of 0.42. According to morpho-kinetical embryo evaluation, our model's PRC of 0.66 significantly improves over existing time-lapse systems for pregnancy prediction, demonstrating better handling of imbalanced clinical data.

The metamodel's calibration metrics (Brier score = 0.20, expected calibration error = 0.06, maximum calibration error = 0.12, Hosmer-Lemeshow test p-value = 0.06) indicate robust reliability in predicting clinical pregnancy outcomes. We validated the model's reproducibility using an independent dataset of 665 treatment cycles, showing close alignment between predicted and actual pregnancy rates (58.9% vs. 59.1%). With the Bayesian method, we proposed a robust framework for integrating historical data with real-time predictions from neural networks, enabling a transition from retrospective to prospective analysis.

Our approach extends beyond conventional embryo selection, incorporating post-analytical phase evaluation in the IVF laboratory. This comprehensive framework enables detailed analysis across different patient subpopulations and time periods, facilitating the identification of systemic issues and IVF protocol optimization. The model's ability to track pregnancy probabilities over time and staff members allows for both outcome prediction and retrospective and prospective assessment of IVF treatment efficacy, providing a data-driven strategy for continuous improvement in assisted reproductive technology.

INTRODUCTION

Despite our extensive knowledge and theories regarding preimplantation embryo development, detailed descriptions of implantation processes, and stringent laboratory protocols, we remain far from fully understanding why success and failure occur in *in vitro* fertilization (IVF) treatments. It's possible that we are thinking in the wrong direction or operating with incorrect parameters. This raises the possibility that a fundamentally different approach to "thinking" and analyzing our work, driven by artificial intelligence (AI), could bring us closer to solving this puzzle.

Although AI systems have found extensive applications in IVF, a detailed algorithm for working with them has not yet been fully developed.[1] It is currently possible to accurately describe the morpho-kinetic changes in individual embryos during *in vitro* culture, predict potential outcomes, and identify the most promising embryos for implantation through such approaches. However, these algorithms cannot always be extrapolated to patients from different clinics in various regions and countries, and their successful application requires additional validation in specific laboratories, taking into account the individual characteristics of the patient population.[2]

In our study, we aimed to integrate laboratory records of individual embryo development, which represent the most accessible and standardized approach for obtaining key quality indicators across laboratories, into a unified neural

a Corresponding Author, Sergei Sergeev, Email: embryossa@gmail.com

network model framework. Our neural network model introduces a novel perspective on the feature selection process essential for predicting the probability of clinical pregnancy with machine learning (ML). Unlike traditional methods that rely on ranking classifications, our model focuses on precise and individualized assessments of key performance indicators (KPIs) with conformal prediction of probabilities from model outputs. It is also important to note that any evaluation system within an IVF laboratory operates as part of a comprehensive quality management framework and should not be considered in isolation. A robust KPI evaluation system is essential to ensure optimal outcomes and continual improvement. A fully functioning quality management and control system must include the process's pre-analytical, analytical, and post-analytical phases. AI models have already found broad application in many of these phases, and the logical outcome of the development of two high-tech fields of modern science is the combination of AI-assisted data analysis and IVF quality management. Currently, the AI algorithms are primarily focused on either the pre-analytical phase (assessing success rates based on initial clinical data and patient treatment history) or the analytical phase — evaluating and selecting individual gametes or embryos based on various parameters.[3,4] Meanwhile, modern AI methods largely overlook the post-analytical phase of data analysis. For its integration, we utilized neural networks in our study because they can identify a broader spectrum of associations than other ML methods, thanks to their ability to recognize highly nonlinear associations among input parameters, and they are less sensitive to data collinearity than all embryo evaluation protocols have.

To fully harness the potential of AI in IVF, it is essential to integrate AI in a single robust pipeline approach, ensuring that each stage of it is effectively managed.[5,6] The concept of the pipeline, drawn from business and engineering fields, represents a well-organized sequence of steps that transform raw inputs into desired outputs, ensuring consistency, quality, and efficiency throughout the process. IVF treatments, like business pipelines, involve multiple stages — each critical to the final outcome, from patient preparation and ovarian stimulation to embryo culture and transfer. Our neural network-based method for predicting pregnancy probabilities adopts this pipeline approach, offering a systematic, data-driven framework that integrates quality assurance and risk management at every phase of IVF treatment.

MATERIALS AND METHODS

DATA COLLECTION AND STUDY DESIGN

We conducted a retrospective analysis of 11,500 IVF treatment cycles from three independent IVF clinics. Each cycle was evaluated using an individual KPIs measurement system.[7] The positive cycle outcome was established as clinical pregnancy that was confirmed by detecting the fetus's heart beating through ultrasound examination 25 days after embryo transfer. All cycles with missing data values were discarded from the study. This dataset was utilized to develop, validate, and test neural network models for predicting clinical pregnancy rates. For that purpose, we split data to train (70%), validate (20%), and test (10%) sets with a stratified random sampling approach.

We adhered to the recommendations set forth by the European Society of Human Reproduction and Embryology (ESHRE) and the Gardner blastocyst grading system for embryo assessment. "good blastocysts" were classified within this framework as those graded 3BB or higher. The KPI analysis covers nine different performance metrics: number of COCs (Cumulus-Oocyte Complexes); fertilization rate; blastocyst rate; TGBDR (Total Good Blastocyst Development Rate); oocyte retrieval rate; KPIScore; MII (mature oocytes); number of blastocysts on day 5; and blastocysts on day 6-7 development.

FEATURE SELECTION

The KPIs we analyzed provide a comprehensive view of embryologists' performance at various stages of IVF, based on the Vienna consensus opinion.[7] The features selected were those identified as having a significant impact on the outcome of IVF treatment according to an independent machine learning (ML) algorithm, XGBoost. These features include: "Age", "Attempt number", "Number of follicles", "Number of oocytes retrieved", "Number of inseminated oocytes", "2PN", "Number of cleaved embryos", "Number of blastocysts", "Number of good blastocysts", "Fertilization rate", "Cleavage rate", "Blastocyst formation rate", "Good blastocyst rate", "Oocyte retrieval rate", "Number of embryos on day 5", "Cryo embryos", "Transfer Day", "Embryos transferred", and ranking sum of the clinical and laboratory indicators - "KPIScore".

Our analysis of individual embryologists' performance revealed no statistically significant differences ($p > 0.05$) in achieving KPIs such as IVF polyspermy rate, ICSI degradation rate, ICSI and IVF fertilization rates, and good blastocyst rate across the selected treatment cycles used for model training.

MODEL VALIDATION

To assess the model's predictive accuracy, we compared the forecasted clinical pregnancy rates against actual outcomes across various time intervals, including quarterly and monthly analyses. For this purpose, we employed a new dataset comprising 665 treatment cycles from the Georgian-German Reproductive Center in Tbilisi, Georgia. A thorough evaluation was performed, contrasting the model's outputs with the observed clinical results.

ETHICAL CONSIDERATIONS

Patient informed consent was not required for this study due to its retrospective nature and the use of fully de-identified embryo development data. No medical interventions were performed on the subjects, and no biological samples were collected from patients for model development, so the

study was entirely non-invasive for both patients and their embryos.

NEURAL NETWORKS DEVELOPMENT

Python 3.11, Scikit-learn 1.4.2, and Sklearn 1.4 were used to implement machine learning models and statistical modeling in DataSpell 2024.2 IDE. The deep learning neural network (DNN) model has been developed and executed in the GPU PyCharm 17.0.10 environment with the Tensorflow 2.15.0 and Keras library 2.14.0. The Kolmogorov-Arnold networks (KAN) were constructed with Pykan package version 0.2.1. Both models were combined with a stacking approach in the final DNN-KAN metamodel. The FTTransformerClassifier was built and evaluated with Mambular 0.2.2 python package in CPU GoogleColab environment to compare our ensemble model with other approaches. All models were calibrated with the Venn-Abers method of conformal prediction with Python package venn-abers 1.4.5 to obtain probabilities of pregnancy achievement from neural network predictions. A comparative analysis of prediction errors was conducted with the area under the receiver operating characteristic curve (AUC), precision-recall curve (PRC), accuracy, F-1 score, specificity (true negative rate), sensitivity (recall), precision (positive prediction value) and Matthew's correlation coefficient (MCC).

To evaluate the performance of the metamodel after calibration, several key quality metrics were utilized: Brier score - quantifies the mean squared difference between predicted probabilities and actual binary outcomes, serving as an indicator of overall model accuracy; expected calibration error (ECE) - the average discrepancy between predicted probabilities and the true outcome frequency; maximum calibration error (MCE) - represents the largest deviation between predicted probabilities and actual outcomes, highlighting the worst-case scenario for model calibration; Hosmer-Lemeshow test - statistical test compares the observed and predicted outcomes in different risk deciles, providing a p-value to assess the goodness of the fit. A non-significant p-value (< 0.05) indicates that the model's predicted probabilities closely align with observed outcomes, reflecting adequate calibration. To evaluate the calibration of our model, we employed the Mean Squared Error (MSE), which allowed us to assess how well-calibrated the model's predictions were, particularly in relation to the real-world data from each clinic. Calibration was further refined by comparing predicted versus observed success rates, ensuring that the model's probabilistic outputs closely aligned with clinical reality.

STATISTICAL ANALYSIS

Descriptive statistics were chosen based on data distribution: for normally distributed quantitative indicators, mean and standard deviation (SD) with 95% confidence intervals (CI) were used, while median and interquartile range (Q1-Q3) were employed for non-normally distributed data, as determined by the Shapiro-Wilk test. For statistical analysis a p-value < 0.05 was used as the significance threshold. Comparison of groups based on quantitative indicators was performed using one-way analysis of variance (ANOVA) for normally distributed data or the Mann–Whitney U test for non-normally distributed data, followed by post hoc comparisons when significant differences were detected. The Chi-square test was used to analyze discrepancies between predicted and actual outcomes of embryo transfers, specifically evaluating the role of staff members in the process.

RESULTS AND DISCUSSION

Kolmogorov-Arnold Networks (KANs) represent a novel approach that offers a compelling alternative to traditional Multi-Layer Perceptrons (MLPs) and Deep Neural Networks (DNNs). While DNNs are rooted in the universal approximation theorem, which guarantees their capability to approximate any continuous function given sufficient depth and complexity, KANs are grounded in the Kolmogorov-Arnold representation theorem.[8] The structural differences between KANs and DNNs significantly affect their performance and interpretability. In DNNs, layers of neurons process data through node-based activation functions, often resulting in highly non-linear transformations that, while powerful, can be challenging to interpret. KANs on the other hand, apply activation functions directly to the interactions between neurons, leading to models that are more interpretable and capable of maintaining or improving accuracy compared to DNNs.

One of the most striking differences in performance between these two architectures is in the distribution of their predictions. KAN models tend to exhibit a wider spread in prediction ranges compared to DNNs, suggesting that KANs may better capture the underlying variability in the data. This characteristic is particularly advantageous in tasks requiring high sensitivity to input variations, such as scientific computing or tasks with complex, small-scale datasets. Given the complementary strengths of DNNs and KANs, integrating these two architectures into a single metamodel presents a sophisticated approach that could significantly enhance predictive accuracy and interpretability. Especially integration is important in complex applications like IVF, where new methods in data analytics with ML and AI are highly conservative.[3]

Training both components of the metamodel on the same dataset ensures that they are exposed to the same underlying patterns and variations in the data. This unified approach allows the metamodel to leverage diverse representations from each network — DNNs providing broad, non-linear approximations and KANs offering detailed, interpretable mappings. The result is a more robust and accurate final prediction, as the metamodel integrates these complementary perspectives. This integration also improves generalization. The DNN-KAN metamodel can reduce the risk of overfitting by balancing the detailed focus of KANs with the broader, high-capacity learning of DNNs. This balance is crucial in domains like IVF outcome prediction, where the dataset may contain highly variable and

sensitive data, and overfitting can lead to unreliable predictions.

Trained on our dataset with balanced class weights, the DNN model achieved an accuracy = 0.70, AUC = 0.74, PRC = 0.64, precision = 0.47, and recall = 0.44. The KAN model had similar performance metrics but with a higher ability to recognize positive pregnancy cases: accuracy = 0.68, AUC = 0.76, PRC = 0.61, precision = 0.61, and recall = 0.62.

The DNN model displays a narrower prediction range, indicating a more conservative approach, with predictions clustering tightly around certain values. The KAN model outperforms the DNN model significantly across all metrics, shows a broader range of predictions, reflecting more flexibility but also potential uncertainty. The KAN model's wider prediction spread and fewer outliers suggest it incorporates a broader range of factors, potentially making it more adaptable but less consistent in certain cases. The DNN model's tighter prediction clustering indicates higher consistency but less adaptability to varied scenarios. In that case DNN model's predictions may be more reliable for clinical settings where consistent decision-making is crucial - QC/QA analysis. The KAN model's range might be beneficial in more nuanced cases where flexibility and a broader view of potential outcomes are needed - individual decisions in single IVF protocol and in troubleshooting methods.

Bothe models significantly (U-Statistic = 230.0, p-value = 0.002) overperformed other ML methods with AUC = 0.64 (CI 0.61 - 0.67) that are reported as models for CPR prediction[9-11] and showed the same AUC metrics with the convolutional neural networks (CNNs) in predicting clinical outcomes with static images (AUC = 0.68–0.71)[12,13] or time-lapse videos (AUC = 0.64–0.67).[14,15]

After fine-tuning the hyperparameters for all evaluated networks, our model's validation loss was 0.56, with a validation accuracy of 69%. This is notably lower than the average validation loss of 0.99 ± 0.15 and higher than the average validation accuracy of 57.54% ± 6.07% achieved by the CNN models for embryo assessment.[16] Following Venn-Abers calibration, both models demonstrated improved performance in terms of probability calibration.

To enhance the power of our approach, we combined both models into one ensemble stacking metamodel. This DNN-KAN model achieved an accuracy of 0.72 (SD = 0.04), AUC of 0.75 (SD = 0.05), PRC of 0.66 (SD = 0.02), precision of 0.70 (SD = 0.02), recall of 0.52 (SD = 0.05), F1 score of 0.60 (SD = 0.05), and maximum MCC of 0.42 on 5-fold cross-validation.

In the next step of our research to identify the best model architecture, we implemented transformer neural networks with our data. Transformer architectures have gained significant attention in various fields, including natural language processing and tabular data analysis, due to their exceptional ability to process sequential data and capture intricate patterns through attention mechanisms. These models are particularly adept at managing complex datasets, such as those encountered in IVF studies, where relationships between variables can be context-dependent.

Given these advantages, we explored using a transformer-based model, specifically the FTTransformerClassifier, to assess its effectiveness compared to our DNN-KAN ensemble. The transformer classifier demonstrated performance metrics comparable to those of our developed metamodel: accuracy = 0.72, AUC = 0.79, PRC = 0.68, precision = 0.63, recall = 0.64, F1 score = 0.63, and MCC = 0.42. This comparison highlights the robustness of our metamodel in handling the specific complexities of our data, emphasizing that while transformers are effective, alternative architectures like our model can perform equally well in certain contexts.

When comparing our metamodel with preanalytical predictive approaches, we observed that its performance metrics align with those reported by other AI-based solutions in IVF, which typically achieve an AUC of 0.62–0.77.[17-20] Furthermore, our metamodel demonstrated comparable performance (U-statistic = 145.0, p-value = 0.471) with analytical time-lapse systems utilizing AI and additional clinical data for CPR and implantation predictions (AUC = 0.72-0.78),[21] as well as with CNN models for static images enhanced with clinical features (AUC = 0.74),[22] and a live birth prediction CNN+MLP model based on multimodal blastocyst evaluation incorporating factors such as maternal age, the day of blastocyst transfer, antral follicle count, retrieved oocyte number, and endometrium thickness (AUC = 0.77, CI = 0.75–0.79).[23]

Despite the significant body of work dedicated to AI in embryo selection, most studies are limited to analyzing and comparing a narrow set of metrics. Traditionally, in the medical field — and in IVF in particular — the performance of a model is often described using the AUC without any calibration of the probabilities. However, it has considerable limitations, especially in the context of imbalanced datasets, which are common in clinical IVF outcome data. In the realm of AI, relying solely on AUC and on raw data of classifications can be misleading and may not always provide accurate insights for subsequent clinical applications. CNN models have gained the most traction in the laboratory phase of embryo selection, primarily by evaluating morpho-kinetic data obtained from time-lapse imaging. However, their utility is contingent on calibrating probabilistic predictions reflecting the true likelihood of success. Poorly calibrated models can lead to misinformed clinical decisions, either overestimating or underestimating the chances of pregnancy. While this approach is widely used, it often overlooks other metrics, such as precision-recall curve (PRC), which may be more suitable for training AI models on imbalanced datasets and offers a deeper evaluation of a model's generalization ability and the reproducibility of its predictions across different IVF centers. Unfortunately, PRC is often neglected and, at best, is only briefly mentioned in the supplementary materials of scientific articles.[15] For instance, one of the most widely used time-lapse assessment models, IDAScore, demonstrates outstanding AUC reporting values as high as 0.95.[24] However, the actual PRC values for this model range from a modest 0.45 to 0.55, indicating a potential disconnect between AUC and the model's real-world predictive perfor-

mance. On the other hand, using static images instead of video sequences has shown better optimization of the precision-recall balance, with AUC values around 0.72,[25] but only a marginal improvement in PRC, which ranges from 0.53 to 0.63. In our metamodel, we achieved mean PRC of 0.66, representing a notable improvement across others and demonstrating that our approach is good enough for handling the imbalanced nature of clinical and laboratory data.

Leveraging the metamodel with different integrated architectures, we performed a comparative analysis of actual and predicted clinical pregnancy rates over various time intervals to assess the likelihood of achieving pregnancy. The conformal prediction calibration of the model output was performed with an odds ratio of 6.01 (SD = 0.65). The main calibration metrics of the model performance were: Brier score = 0.20; ECE = 0.06; MCE = 0.12; Hosmer-Lemeshow test statistic = 15.25 with p-value = 0.06. These results indicate that the final model exhibits a robust calibration with a well-balanced Brier Score and a low ECE, suggesting accurate probability predictions and signifying that the predicted probabilities are closely aligned with the actual outcomes. The low level of MCE highlights improved consistency across different patient groups, and the Hosmer-Lemeshow test shows a non-significant p-value, reflecting that the predicted probabilities closely match the observed outcomes. According to that, we can conclude that Venn-Abers conformal prediction provides reliable uncertainty estimates, which are critical for clinical decision-making in IVF treatments.

As with all AI systems, one of the key limitations of the developed metamodel is the necessity for specific calibration to individual IVF clinics. Without this adaptation, the model's predictive power is diminished due to the inherent variability in patient demographics, treatment protocols, and clinical practices across different IVF centers. For instance, when the non-calibrated model was applied across nine clinics in different cities in Russia and Kazakhstan, the AUC scores ranged from 0.55 to 0.73, with MCE = 0.89 and MSE = 0.19. This variability highlights the model's inconsistent ability to distinguish between positive and negative pregnancy outcomes without the step of propria calibration. Such variations may result in biased predictions and reduced clinical utility in settings where the model has not been fine-tuned to account for specific clinical and population characteristics.

Further optimization of the site-specific calibration process could improve predictive accuracy even more. However, the current performance on well-calibrated data with MCE = 0.12 and MSE = 0.17 already demonstrates the effectiveness of combining DNN and KAN architectures within a metamodel framework.

This approach enabled us to set a lower threshold for the probability of clinical pregnancy occurrence for each year of operation. Our analysis revealed a statistically significant difference ($p < 0.05$) in the likelihood of clinical pregnancy between the 2021-2022 years cohort of patients and that of the 2023 year, with a decline in the likelihood of clinical pregnancy ranging from 10% at the beginning of the year to 20% by the third quarter of 2023 year. But no significant differences were found for the following KPIs in that period: fertilization rate (F-statistic = 3.47, p-value = 0.112); blastocyst rate (F-statistic = 0.0083, p-value = 0.930); TGBDR (F-statistic = 0.633, p-value = 0.457); MII rate (F-statistic = 0.133, p-value = 0.727); follicular oocyte index (F-statistic = 0.045, p-value = 0.839). Neither significant difference was observed with 2024-year data when we don't have any serious shifts of CPR: fertilization rate (F-statistic = 1.42, p-value = 0.205); blastocyst rate (F-statistic = 0.58, p-value = 0.769); TGBDR (F-statistic = 1.17, p-value = 0.326); MII rate (F-statistic = 1.42, p-value = 0.207); follicular oocyte index (F-statistic = 1.93, p-value = 0.072). The probabilities calculated theoretically using the metamodel were consistent with the actual pregnancy outcomes during these periods. This alignment between our KPIs calculation and the metamodel's predictions suggests that the observed decline in clinical pregnancy rates from the third quarter of the 2023 year is not attributable to the quality of patient preparation or laboratory procedures but rather to the initial clinical data of the patients.

It is remarkable that despite the variations in embryo transfer outcomes and patient groups, our laboratory KPIs were at a benchmark level according to the Vienna consensus[7]: mean fertilization rate = 0.83 (CI = 0.70-0.96); mean blastocyst rate = 0.52 (CI = 0.48-0.56); mean TGBDR = 0.43 (CI = 0.37-0.50). This provides a good data source for understanding the process and concluding that all noted changes in CPR across different periods do not originate from the embryology laboratory in this case.

Clinical and laboratory components are intricately intertwined in the IVF process, each playing a vital role in determining treatment success. This interdependence necessitates a comprehensive and collaborative approach to quality management, where both of them are aligned to achieve optimal patient outcomes.[26] Our research aimed to address this need by comparing outcomes between staff members in cases where our metamodel predictions did not match actual results. Differentiating between patients treated by various doctors within a clinic often presents challenges, making it difficult to define clear competency boundaries in achieving quality targets. We compared actual CPRs achieved by individual reproductive specialists throughout 2023, using external audit programs across four different IVF centers. This analysis identified three doctors (No. 1, No. 2, and No. 3) with actual CPRs of 34.0%, 42.5%, and 40.5%, respectively, which exceeded their corresponding theoretical thresholds of 33.60%, 33.33%, and 38.01% ($p > 0.05$). Conversely, three other doctors (No. 4, No. 5, and No. 6) exhibited significantly lower CPRs (9.60%, 24.82%, and 16.01%) compared to the model's predicted probabilities of 33.53%, 33.33%, and 37.25% for their patient groups ($p < 0.01$). The median values for fertilization and blastocyst development rates exceeded the Vienna consensus benchmark for all conducted cycles. Based on these results, it can be concluded that doctors No. 4, No. 5, and No. 6 currently do not possess the necessary competency to work independently. Procedures conducted by these staff members require close supervision and mentoring from more experi-

enced colleagues, such as doctors No. 1, No. 2, and No. 3. So, our pipeline is designed to offer a holistic approach that integrates clinical and laboratory data while providing individualized assessments of each participant's performance in the process. This is achieved by evaluating the quality indicators of individual staff members and their impact on embryo transfer outcomes. Identifying areas for improvement in clinical procedures and laboratory protocols is essential, ensuring that every aspect of the IVF process is optimized for success.

To compare outcomes where our metamodel's predictions did not align with actual clinical results, we analyzed 64 complex cases from 665 embryo transfer results obtained in 2024. The Chi-square analysis yielded a p-value of 0.633 for embryologists, indicating no statistically significant differences. Similarly, the analysis produced a p-value of 0.588 for physicians, also showing no statistically significant differences. Interestingly, the p-value for embryologists was higher than that for physicians. Although these findings do not reach statistical significance, they may suggest a trend where the physician performing the embryo transfer has a more substantial impact on the outcome than the embryologist responsible for thawing and preparing the embryos for transfer. However, this observation requires careful interpretation and should not diminish the importance of standardizing laboratory procedures and thoroughly analyzing laboratory performance metrics.

Our data analysis further demonstrated the high accuracy of our developed model in predicting CPR, particularly evident when examining quarterly and monthly trends. For instance, in the second quarter of the 2024 year, following the integration of a comprehensive quality control system, the predicted pregnancy rate was 58.9%, closely matching the actual rate of 59.1%, with a minimal difference of just 0.2 percentage points. The time series analysis of pregnancy probabilities highlighted significant variability ($p < 0.05$) in individual case-level predictions, reflecting the complexity of predicting IVF success. KPIs analysis revealed no significant differences ($p > 0.05$) in pairwise comparisons between all embryologists during that period. However, when data were aggregated on a monthly or quarterly basis, a clear trend of improved prediction accuracy over time emerged. This trend was especially noticeable from late 2023 into 2024, where the predicted and actual rates showed near-perfect alignment. Specifically, the mean absolute error (MAE) between predicted and actual rates decreased from 0.15 in the first quarter of the 2022 year to 0.03 in the second quarter, indicating a statistically significant improvement ($p < 0.05$, paired t-test). This improvement can be attributed to several factors, including the increased volume of training data and iterative enhancements to the model's algorithm.

Despite the overall high accuracy, there were instances where the model's predictions deviated from actual outcomes. These deviations can be attributed to factors such as clinical control changes — where improvements in stimulation protocols and embryo transfer techniques might influence outcomes that the model cannot immediately account for — and external variables such as seasonal fluctuations, changes in patient demographics, or other medical factors that may impact results without being fully captured by the model.[27] For example, in the second quarter of 2022 year, the model predicted a pregnancy rate of 14.3%, whereas the actual rate was 43.3%, marking the largest observed discrepancy in our dataset. This deviation coincided with a period of laboratory reconstruction and suboptimal functionality, leading us to exclude this period from further analysis.

In extending our analysis, we incorporated the Bayesian method to transition from retrospective to prospective prediction in IVF outcomes. The Bayesian approach provides a rigorous framework that combines historical data with real-time predictions, enhancing the accuracy of our forecasting models. This method is particularly advantageous in IVF, where integrating clinical experience with machine learning outputs can significantly improve the reliability of predictions.

By treating the neural network's predicted probabilities as new observations, we calculated the posterior distribution, which merges past data with current predictions: a predicted success rate of approximately 58% for the next cycle, with a confidence interval ranging from 55% to 61%. This updated distribution offers a more refined estimate of the expected success rate, now inclusive of both historical trends and real-time insights, and serves as a prospective benchmark for our own laboratory, guiding our clinical decision-making and providing a clear reference point for assessing the effectiveness of our IVF protocols in upcoming cycles.

With our pipeline, we performed KPI analysis for different patient subpopulations undergoing IVF treatment in our center across three groups: China, Georgia, and Israel, according to our metamodel prediction distribution. First, we compared the most promising programs with the majority of oocyte donor cycles of China patients and for bad prognosis patients from Georgia. Statistically significant differences between them were observed for the following indicators: blastocyst rate ($p = 0.004$), TGBDR ($p = 0.009$), FOI ($p = 0.004$). No statistically significant differences were found for fertilization rate ($p = 0.51$) and MII rate ($p = 0.52$), which proves the invariance in treatment approaches and laboratory parameters for all patient groups. Statistically significant differences between China and Israel groups were observed for MII rate ($p < 0.001$) and FOI ($p < 0.001$) due to different ovarian stimulation approaches in these patients. The same differences were found for Georgia and Israel patients with consistent laboratory performance indicators: fertilization rate ($p = 0.69$), blastocyst rate ($p = 0.41$), TGBDR ($p = 0.83$). We achieved predicted CPR: China – 63.0%, Israel – 46.1%, Georgia – 46.5% which is not more than 2% differ with the actual pregnancy rate in that groups. Similar results were obtained from patients' data in Russia with a mean predicted CPR of 36% (CI = 29-37%), which fully aligned with actual clinical data.

Utilizing our pipeline approach, we performed a comprehensive comparison of distributions for various key factors between problem and non-problem cases in IVF cycles.

We observed that successful IVF cycles generally tend to have higher numbers of COCs and mature oocytes, indicating better ovarian response and oocyte quality; higher blastocyst formation rates are associated with non-problem cycles; fertilization rates show less pronounced differences, suggesting that other factors may be more critical in determining cycle success. These findings provide valuable insights into the factors differentiating successful and unsuccessful IVF cycles, potentially guiding improvements in clinical protocols and patient management strategies.

CONCLUSION

This research contributes to the growing body of work aimed at integrating AI technologies into IVF treatment, where predictive models hold the potential to improve patient outcomes by personalizing treatment strategies. The combination of DNN and KAN architectures represents an innovative approach within AI-assisted IVF protocols, aligning with broader trends in medicine where ML and neural networks are being used to optimize complex decision-making processes. As IVF technology continues to evolve, incorporating AI models tailored to specific clinical environments could significantly enhance the precision of fertility treatments, contributing to higher success rates and more individualized care. This study adds to these advancements by addressing the need for an adaptable, robust AI pipeline capable of handling the inherent variability of IVF patients and clinics.

A notable innovation in our work is the integration of two fundamentally different neural network architectures, DNN and KAN, combined with the Venn-Abers approach for calibrated prediction. By merging these distinct learning methodologies, we developed a powerful metamodel with AUC = 0.75, PRC = 0.66, and MCE = 0.12, transforming data analytics in our quality management system. Unlike traditional time-lapse imaging, which often carries high uncertainty in predictions, our approach extends predictive power beyond embryo selection, encompassing the entire IVF process, including the post-analytical phase. This comprehensive framework allows for enhanced quality assurance and risk minimization at every stage of treatment. By incorporating a wide range of laboratory and clinical parameters, our model evaluates the overall effectiveness of IVF protocols and identifies systemic issues that may contribute to unsuccessful outcomes.

The ongoing refinement of this model, along with continuous improvements in clinical and laboratory processes, promises to further enhance the success of IVF treatments. By providing a comprehensive, data-driven framework for analyzing and optimizing IVF protocols, this approach can lead to more effective infertility treatments. Future research should focus on expanding the dataset to include more diverse patient demographics and clinical scenarios, incorporating additional factors that may influence IVF success rates, developing real-time predictive capabilities for immediate protocol adjustments, and exploring the potential for personalized treatment optimization based on individual patient profiles and predicted outcomes with site-specific calibration. Without it, the model's predictive accuracy diminishes across clinics. This underscores the need for site-specific adaptation to improve generalizability. Additionally, the model's reliance on retrospective data limits its ability to capture real-time changes in clinical practice or patient demographics. As IVF protocols evolve and new technologies emerge, there is a risk that the model's predictions may become less accurate unless periodically re-calibrated and updated with more recent data.

AUTHORS' CONTRIBUTION

Conceptualization: [Sergei Sergeev]; Methodology: [Sergei Sergeev]; Data curation: [Lasha Nadirashvili]; Writing - original draft preparation: [Sergei Sergeev], [Iuliia Diakova]; Writing - review and editing: [Iuliia Diakova]; Resources: [Sergei Sergeev]; Validation: [Iuliia Diakova], [Lasha Nadirashvili]; Supervision: [Iuliia Diakova]

COMPETING OF INTEREST – COPE

No competing interests were disclosed

INFORMED CONSENT STATEMENT

All authors and institutions have confirmed this manuscript for publication

DATA AVAILABILITY STATEMENT

Data and code are available at: https://github.com/embryossa/KAN-in-IVF

Submitted: September 16, 2024 CST, Accepted: October 14, 2024 CST

REFERENCES

1. Jiang VS, Bormann CL. Artificial intelligence in the in vitro fertilization laboratory: a review of advancements over the last decade. *Fertil Steril*. 2023;120(1):17-23. doi:10.1016/j.fertnstert.2023.05.149

2. Kromp F, Wagner R, Balaban B, et al. An annotated human blastocyst dataset to benchmark deep learning architectures for in vitro fertilization. *Sci Data*. 2023;10(1):271. doi:10.1038/s41597-023-02182-3

3. Güell E. Criteria for implementing artificial intelligence systems in reproductive medicine. *Clin Exp Reprod Med*. 2024;51(1):1-12. doi:10.5653/cerm.2023.06009

4. Yang HY, Leahy BD, Jang WD, et al. BlastAssist: a deep learning pipeline to measure interpretable features of human embryos. *Hum Reprod*. 2024;39(4):698-708. doi:10.1093/humrep/deae024

5. Simopoulou M, Sfakianoudis K, Maziotis E, et al. Are computational applications the "crystal ball" in the IVF laboratory? The evolution from mathematics to artificial intelligence. *J Assist Reprod Genet*. 2018;35(9):1545-1557. doi:10.1007/s10815-018-1266-6

6. Fernandez EI, Ferreira AS, Cecílio MHM, et al. Artificial intelligence in the IVF laboratory: overview through the application of different types of algorithms for the classification of reproductive data. *J Assist Reprod Genet*. 2020;37(10):2359-2376. doi:10.1007/s10815-020-01881-9

7. ESHRE Special Interest Group of Embryology and Alpha Scientists in Reproductive Medicine. The Vienna consensus: report of an expert meeting on the development of ART laboratory performance indicators. *Reprod Biomed Online*. 2017;35(5):494-510. doi:10.1016/j.rbmo.2017.06.015

8. Schmidt-Hieber J. The Kolmogorov-Arnold representation theorem revisited. *Neural Netw*. 2021;137:119-126. doi:10.1016/j.neunet.2021.01.020

9. Li L, Cui X, Yang J, Wu X, Zhao G. Using feature optimization and LightGBM algorithm to predict the clinical pregnancy outcomes after in vitro fertilization. *Front Endocrinol (Lausanne)*. 2023;14:1305473. doi:10.3389/fendo.2023.1305473

10. Sarais V, Reschini M, Busnelli A, Biancardi R, Paffoni A, Somigliana E. Predicting the success of IVF: external validation of the van Loendersloot's model. *Hum Reprod*. 2016;31(6):1245-1252. doi:10.1093/humrep/dew069

11. Bori L, Meseguer F, Valera MA, Galan A, Remohi J, Meseguer M. The higher the score, the better the clinical outcome: retrospective evaluation of automatic embryo grading as a support tool for embryo selection in IVF laboratories. *Hum Reprod*. 2022;37(6):1148-1160. doi:10.1093/humrep/deac066

12. VerMilyea M, Hall JMM, Diakiw SM, et al. Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. *Hum Reprod*. 2020;35(4):770-784. doi:10.1093/humrep/deaa013

13. Enatsu N, Miyatsuka I, An LM, et al. A novel system based on artificial intelligence for predicting blastocyst viability and visualizing the explanation. *Reprod Med Biol*. 2022;21(1):e12443. doi:10.1002/rmb2.12443

14. Kragh MF, Rimestad J, Lassen JT, Berntsen J, Karstoft H. Predicting Embryo Viability Based on Self-Supervised Alignment of Time-Lapse Videos. *IEEE Trans Med Imaging*. 2022;41(2):465-475. doi:10.1109/TMI.2021.3116986

15. Berntsen J, Rimestad J, Lassen JT, Tran D, Kragh MF. Robust and generalizable embryo selection based on artificial intelligence and time-lapse image sequences. *PLoS One*. 2022;17(2):e0262661. doi:10.1371/journal.pone.0262661

16. Thirumalaraju P, Kanakasabapathy MK, Bormann CL, et al. Evaluation of deep convolutional neural networks in classifying human embryo images based on their morphological quality. *Heliyon*. 2021;7(2):e06298. doi:10.1016/j.heliyon.2021.e06298

17. Benchaib M, Labrune E, Giscard d'Estaing S, Salle B, Lornage J. Shallow artificial networks with morphokinetic time-lapse parameters coupled to ART data allow to predict live birth. *Reprod Med Biol*. 2022;21(1):e12486. doi:10.1002/rmb2.12486

18. Goyal A, Kuchana M, Ayyagari KPR. Machine learning predicts live-birth occurrence before in-vitro fertilization treatment. *Sci Rep*. 2020;10(1):20925. doi:10.1038/s41598-020-76928-z

19. Erlich I, Ben-Meir A, Har-Vardi I, et al. Pseudo contrastive labeling for predicting IVF embryo developmental potential. *Sci Rep*. 2022;12(1):2488. doi:10.1038/s41598-022-06336-y

20. Loewke K, Cho JH, Brumar CD, et al. Characterization of an artificial intelligence model for ranking static images of blastocyst stage embryos. *Fertil Steril*. 2022;117(3):528-535. doi:10.1016/j.fertnstert.2021.11.022

21. Lee CI, Huang CC, Lee TH, et al. Associations between the artificial intelligence scoring system and live birth outcomes in preimplantation genetic testing for aneuploidy cycles. *Reprod Biol Endocrinol*. 2024;22(1):12. doi:10.1186/s12958-024-01185-y

22. Miyagi Y, Habara T, Hirata R, Hayashi N. Predicting a live birth by artificial intelligence incorporating both the blastocyst image and conventional embryo evaluation parameters. *Artif Intell Med Imaging*. 2020;1(3):94-107. doi:10.35711/aimi.v1.i3.94

23. Liu H, Zhang Z, Gu Y, et al. Development and evaluation of a live birth prediction model for evaluating human blastocysts from a retrospective study. *eLife*. 2023;12:e83662. doi:10.7554/eLife.83662

24. Berntsen J, Rimestad J, Lassen JT, Tran D, Kragh MF. Robust and generalizable embryo selection based on artificial intelligence and time-lapse image sequences. *PLoS One*. 2022;17(2):e0262661. doi:10.1371/journal.pone.0262661

25. Kim HM, Kang H, Lee C, et al. Evaluation of the Clinical Efficacy and Trust in AI-Assisted Embryo Ranking: Survey-Based Prospective Study. *J Med Internet Res*. 2024;26:e52637. doi:10.2196/52637

26. Mittal M, Supramaniam PR, Lim LN, Hamoda H, Savvas M, Narvekar N. Is the clinician an independent variable in embryo transfer outcomes under standardized direct and indirect supervision? A 5-year observational cohort study. *GMS J Med Educ*. 2019;36(1):Doc7. doi:10.3205/zma001215

27. Singh A, Joseph T, Karuppusami R, Kunjummen AT, Kamath MS, Mangalaraj AM. Seasonal Influence on Assisted Reproductive Technology Outcomes: A Retrospective Analysis of 1409 Cycles. *J Hum Reprod Sci*. 2021;14(3):293-299. doi:10.4103/jhrs.jhrs_39_21